LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-422243

Task 1.5 Genomic Shift and Drift Trends of Emerging Pathogens

M. Borucki

January 8, 2010

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Task 1.5 Genomic Shift and Drift Trends of Emerging Pathogens

The Lawrence Livermore National Laboratory (LLNL) Bioinformatics group has recently taken on a role in DTRA's Transformation Medical Technologies Initiative (TMTI). The high-level goal of TMTI is to accelerate the development of broad-spectrum countermeasures. To achieve those goals, TMTI has a near term need to conduct analyses of genomic shift and drift trends of emerging pathogens, with a focused eye on select agent pathogens, as well as antibiotic and virulence markers.

Executive Summary

Most emerging human pathogens are zoonotic viruses with a genome composed of RNA. The high mutation rate of the replication enzymes of RNA viruses contributes to sequence drift and provides one mechanism for these viruses to adapt to diverse hosts (interspecies transmission events) and cause new human and zoonotic diseases. Additionally, new viral pathogens frequently emerge due to genetic shift (recombination and segment reassortment) which allows for dramatic genotypic and phenotypic changes to occur rapidly. Bacterial pathogens also evolve via genetic drift and shift, although sequence drift generally occurs at a much slower rate for bacteria as compared to RNA viruses. However, genetic shift such as lateral gene transfer and inter- and intragenomic recombination enables bacteria to rapidly acquire new mechanisms of survival and antibiotic resistance. New technologies such as rapid whole genome sequencing of bacterial genomes, ultra-deep sequencing of RNA virus populations, metagenomic studies of environments rich in antibiotic resistance genes, and the use of microarrays for the detection and characterization of emerging pathogens provide mechanisms to address the challenges posed by the rapid emergence of pathogens. Bioinformatic algorithms that enable efficient analysis of the massive amounts of data generated by these technologies as well computational modeling of protein structures and evolutionary processes need to be developed to allow the technology to fulfill its potential.

Introduction to virus evolution

Two mechanisms by which viruses evolve and emerge are genetic drift and genetic shift. In this report, “genetic drift” refers to point mutations (single nucleotide substitutions, insertions or deletions) introduced during genome replication due to polymerase infidelity (Simmonds, 2004). Because all replication enzymes have some rate of error and even a single nucleotide change may result in a change in the amino acid sequence and thus affect the function of a protein, genetic drift is an important mechanism of viral evolution.

The mutation rate of most RNA viruses is about 1,000 times faster than that of most DNA viruses. It is estimated at 10^{-3}-10^{-5} substitutions per nucleotide copied; that is, these viruses insert the incorrect nucleotide in about 1 out of 1,000 bases to 1 out of 100, 000 bases every time the genome is replicated. The small genome size, short generation times, and large population sizes characteristic of RNA viruses allows these high error rates to contribute to the genetic diversity of the population without leading to extinction (Elena et al, 2006). For this reason a single RNA virus population can consist of individual viruses with slightly different genomes, sometimes referred to as a mutant swarm or viral quasispecies (Domingo, 2007). It is generally understood that the high mutation rate of RNA viruses increases the ability of these viruses to adapt to diverse hosts (interspecies transmission events) and cause new human and zoonotic diseases. However, very little is known about the particular mutations that enable interspecies transmission events to occur.

Viruses may also evolve by a more dramatic mechanism referred to as genetic shift. Genetic shift occurs when two different but closely related viruses exchange an entire fragment of their genome during the process of replication; this is known as recombination. Because this exchange can involve entire genes, genetic shift can drastically change the phenotype of the recombinant virus. Recombination occurs more commonly in viruses that have a positive-sense polarity to their genome versus those with a negative-sense RNA genome. Genetic shift also occurs when two closely related segmented viruses infect the same cell and exchange entire segments of their genome. Two examples of viruses that have recently emerged as a result of genetic shift are the SARS virus (a recombinant virus) and H1N1 “swine” influenza virus (a reassortant virus).

General trends in viral emergence

A vast majority of recently emerged human pathogens are viruses, and two-thirds of those are RNA viruses (Woolhouse, 2006). Most human pathogens are zoonotic (have an animal reservoir) as do all of the HHS Select Agent pathogens with the exception of Variola (Figure 1, Table 1). The evolutionary capacity of a virus to evolve and infect new species depends on the composition of the genome (RNA versus DNA), polarity of the genome (recombination is more common is positive-sense RNA viruses), and number of genome segments (multi-segmented viruses can reassort genome segments) (Parrish et al. 2008).

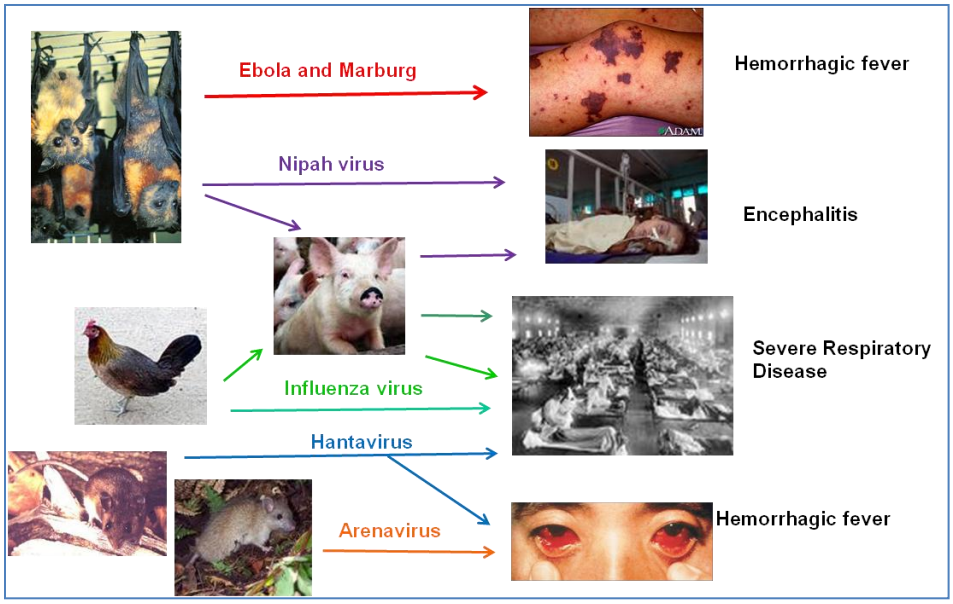

Figure 1. NIAID category A viruses which have an animal reservoir.

In order for a zoonotic virus to infect humans, the virus must have proteins on its surface that allow it to bind to and enter human cells. Therefore, many emergent viruses use cellular receptors that are conserved across a variety of species. Additionally, the distribution of the receptor on various host tissues influences the course of the infection, for example, a receptor that is present on many tissue types may allow the infection to become systemic rather than remaining localized. Although computational modeling of protein binding is still in the formative stages, *in silico* modeling of the viral ligand (attachment protein) of a novel virus may someday allow identification of the host receptor(s) and subsequent prediction of host range and tissue tropisms.

Once a virus has attained the ability to infect a new cell type, the virus must then be able to effectively replicate in that cell type, evade the immune response, and be transmitted efficiently by the new host. Therefore, emergence of a new pathogen is thought to occur in a step-wise process (Parrish et al., 2008). For example, it is believed that SARS virus was the result of two different mechanisms of virus evolution; genome recombination resulted in the replacement of the spike protein gene of a bat coronavirus with that of a human coronavirus and allowed the recombinant virus to infect human cells. Subsequent generation of point mutations allowed the new recombinant virus to be transmitted efficiently among humans (Li et al., 2007).

Viral Family (Virus)	Genome	No. of Segments	Mutation Rate	Recombine	Reassort	Animal Reservoir	References
Filoviridae (Marburg, Ebola)	RNA, ss -	1	10(-4)	Yes	NA	Bats	(Hanada et al, 2004; Wittmann et al., 2007)
Arenaviridae (Lassa, Junin, Machupo, etc)	RNA, ss +/-	2	10(-3)	Yes	Yes	Rodents	(Hanada et al, 2004; Emonet et al., 2009)
Bunyaviridae (Rift Valley fever, Crimean-Congo HF)	RNA, ss -, +/-	3	10(-3)	Yes	Frequent	Arthropods Rodents Livestock	(Hanada et al, 2004; Deyde et al., 2006)
Flaviviridae (Tick-borne encephalitis)	RNA, ss +	1	10(-4)	Rare	NA	Arthropods Primates Birds	(Hanada et al, 2004; Gubler, 2007)
Paramyxo-viridae (Hendra, Nipah)	RNA, ss -	1	10(-3)	Rare, Absent	NA	Bats	(Chare et al., 2003)
Togaviridae: Alphaviruses (Eastern and Venezuelan equine encephalitis)	RNA, ss +	1	10(-4)	Yes, in vitro and in vivo (Infrequent)	NA	Arthropods Birds Rodents	(Domingo, 2007)
Orthomyxo-viridae (Influenza)	RNA, ss -	8	10(-3)	Rare, Absent	Frequent	Birds Swine	(Domingo, 2007; Boni et al., 2008)
Coronaviridae (SARS)	RNA, ss +	1	10(-3)	Frequent	NA	Bats Palm civets	(Lai et al., 2007)
Poxviridae (Smallpox, Monkey pox)	DNA, ds	1	10(-6)	Likely	NA	Primates Rodents (Monkey pox)	(Babkin et al., 2006, 2008; Esposito et al., 2006)
Herpesviridae (Herpes B)	DNA, ds	1	10(-8)	Rare	NA	Primates	(Sharp, 2002; Vasilakis et al., 2009)

Table 1. Characteristics of viral families represented in the HHS list of select agent viruses.
NA= Not applicable; ss= single-stranded, ds= double-stranded; + = positive-sense polarity, - = negative-sense polarity

Virus transmission by arthropods such as mosquitoes or ticks provides an additional mechanism for virus evolution and dissemination. These arthropod-borne viruses ("arboviruses") have a great capacity for emergence due to the fact that most have RNA genomes and infection of the arthropod vectors may increase the potential for virus dispersal. Additionally, the life-long, chronic infection of an arthropod with a virus may provide an opportunity for the host to become "superinfected" with a second, closely related virus, thus providing an opportunity for genetic shift to occur (Reese et al., 2008). However, genetic drift occurs at a slightly lower rate for most arboviruses. In fact, the mutation rate of arboviruses is generally ten times less than other RNA viruses likely due to the constraints imposed by replication in two different types of hosts, invertebrates (arthropods) and vertebrates (mammals, birds) (Vasilakis et al., 2009).

Finally, the chance of a zoonotic virus mutating to infect humans also depends on the amount and type of contact between the animal species and humans. The density of the animal host may also affect virus evolution (Moya et al., 2004). For example, animals such as bats which live in large, dense populations may harbor many different but closely related viruses, and the close contact among animals in these communities provide ample opportunity for the host to become infected with two different viruses, thus providing an opportunity for genetic shift to occur (Pourrut et al., 2007). Examples of this include the role of pig farms in the emergence of viruses such as influenza and Nipah virus, the role of fruit bat colonies in the introduction of diverse *Ebolavirus* strains into the human population, and the role of live animal markets in the transmission of H5N1 avian influenza and SARS virus.

The Role of Recombination in the Emergence of Novel Coronaviruses

Multiple new coronaviruses, including the SARS virus, have been identified in human samples just within the last five years, demonstrating the potential of coronaviruses as emergent human pathogens. An important feature of coronavirus evolution is the frequency at which the viral genomic RNA recombines with the genomes of other viruses (calculated to be as high 25% among closely related viruses) (Lai et al., 2007). Genetic analysis of the SARS virus genome indicates this novel pathogen was the result of two different mechanisms of virus evolution; genome recombination of the spike protein gene, followed by the generation of point mutations. The most likely reservoir of the ancestor to the SARS virus are coronaviruses isolated from bats. However, bat coronaviruses have some genetic differences in the spike protein as compared to SARS. The spike protein is located on the exterior of the virus particle and is responsible for host cell binding and invasion. Therefore, the host range of a specific coronavirus is largely determined by its spike protein (Li et al., 2007). It is believed that a bat coronavirus and a human cold coronavirus co-infected a host (such as a civet cat) and that recombination of the two viral genomes created a SARS-like virus which then evolved additional single-nucleotide mutations that increased its ability to infect human cells (Hanada et al., 2004).

Interestingly, recombination may also occur between coronaviruses and other families of viruses. The presence of the HE gene in coronaviruses is believed to have resulted from a recombination event between influenza C and a subgroup 2 coronavirus. Again, this speaks to the potential of coronaviruses to undergo dramatic evolutionary steps.

A white paper has been submitted for the use of new sequencing technologies combined with advanced computational analysis to identify the role of genetic drift (ultra-deep sequence analysis of quasispecies) and shift (recombination frequency and hot spots) in adaptation of a zoonotic coronavirus to growth in human cells.

The Role of Reassortment in Bunyavirus Emergence

Rift Valley fever is a viral disease that is present in much of Africa and has recently spread to Egypt and the Arabian Peninsula. Although Rift Valley fever is primarily a disease of livestock, it can spread to humans and cause large outbreaks of potentially fatal disease. Rift Valley fever virus (RVFV) causes epidemics of severe disease in cattle, sheep, and goats that often spread to adjacent human populations via bites from infected mosquitoes or exposure to contaminated animal tissues and fluids. In humans, RVFV infection usually results in an acute, self-limiting flu-like illness, however a small percentage of cases progress to more severe disease such as viral hemorrhagic fever and encephalitis.

RVFV belongs to the family *Bunyaviridae*, a group of viruses with negative-sense RNA genomes that consist of three segments (Table 1). Many members of this family are important human and veterinary pathogens, and most members of this family are transmitted by arthropods, with the exception being members of the genus *Hantavirus* which are transmitted by rodents. Viruses transmitted to vertebrates (i.e. mammals, birds) by mosquitoes generally have relatively stable genomes as the need to infect two divergent hosts appears to constrain genetic drift (Vasilakis et al., 2009), however genetic shift via recombination or segment reassortment may occur in arthropod-borne viruses.

RVFV is transmitted primarily by mosquitoes, and this serves as a mechanism for rapid evolution via genetic shift. In this case, a mosquito becomes infected by two different strains of RVFV and the virus exchange genomic segments in dually-infected mosquito cells. Novel viruses may also emerge due to segment reassortment between different species of bunyavirus (Briese et al., 2007; Gerrard et al., 2004; Borucki et al., 1999).

A recent study by Bird et al. (2007) which examined the genetic diversity of geographically, temporally, and phenotypically diverse RVFV strains, found that the genomes sequenced were highly conserved and as well as evidence for segment reassortment. This study also identified amino acid differences between lineages of RVFV that may affect the pathogenicity of the virus

in a rodent model. Communications between members of our team and CDC's Stuart Nichol (the senior author of the study) have identified a common interest for more in-depth analysis of these strains including the development of computational tools for RVFV virulence prediction studies.

Analysis of Genetic Shift in Influenza A

Although seasonal (H3N2) influenza A is not a select agent, the Influenza Genome Sequencing Project initiated the first large scale whole genome viral sequencing effort to focus on surveying the broad evolutionary landscape of the virus (Ghedin et al. 2005), as opposed to past efforts that focused on sequencing isolates of unique medical interest (e.g. highly pathogenic strains). This effort, therefore, gives a useful case study for identifying what insights into viral evolution associated with emerging threats might be gained should similar efforts be undertaken for other viruses of interest. Perhaps, the most strikingly unique early result from the whole influenza genome sequencing work was the observation that genetic reassortment within H3N2 strains plays a prominent role in the virus' evolution (Holmes et al. 2005). Ongoing sequencing efforts have since made clear that reassortment as a form of genetic shift occurs across many strains. For example, reassortment within multiple H1N1 strains contributed to the emergence of the recent H1N1 pandemic strain (Influenza investigation team, 2009). Similarly, reassortment among influenza strains that normally infect different hosts has been observed in isolated cases of human infection (Olsen et al., 2006). Despite the increasing appreciation for the role reassortment plays in influenza evolution, detecting reassortment events from whole genome sequencing remains a non-trivial task and can require extensive phylogenetic analysis.

The work to detect reassortment events is part of a broader effort to develop computational methods that incorporate genetic shift events into new models of evolution. These more complex evolutionary histories can be described with phylogenetic networks rather than phylogenetic trees (Huson and Bryant, 2006). Two examples are split networks, which identify inconsistencies between multiple phylogenetic tree representations, and can be used to infer genetic shift. Reticulate networks are a second type of phylogenetic network, which reconstruct an ancestral history that includes nodes with edges from multiple ancestral strains and can contain multiple trees representing the presence of an ancestral genetic shift event (Nakhleh et al, 2005). A schemetatic taken from Houson and Bryant in Figure 2 shows a wide array of methods for inferring genetic shift and the typical data sources used as input.

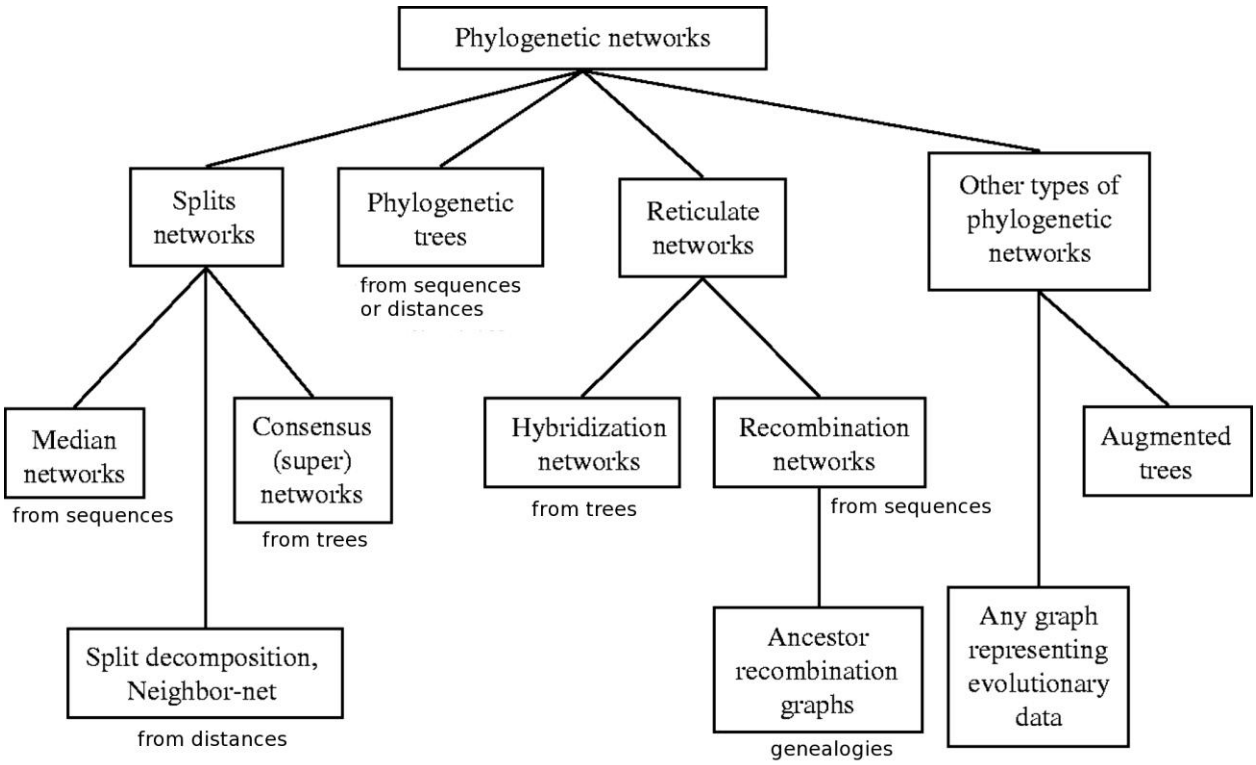

Figure 2. Schematic of different methods for inferring genetic shift.

Recent bioinformatics work in this area is showing some potential to develop automated computational pipelines that make use of large scale genome sequence data to identify reassortment in influenza A (Rabadan et al. 2008, Nagarajan and Kingsford 2008). This remains an active area of research, and many methods, which make use of phylogenetic inference benefit from large computational resources. One particularly intriguing approach, which requires relatively limited computational resources, compares segment pairs from two influenza isolates independently, rather than through the reconstruction of a phylogeny from multiple sequences simultaneously. The method assumes that when reassortment is absent, the segment pairs from the two isolates should both share the same most recent common ancestor. This can be tested by measuring the genetic distance for each pair of segments for two isolates. For example, segment 1 from isolate A and isolate B are compared and given a genetic distance g_1. Assuming a constant rate of evolution, when segment 2 from isolate A and isolate B are compared, the genetic distance g_2, should be roughly similar to g_1. An expected distribution of the differences between the genetic distances g_1 and g_2 from segment pairs can be calculated to identify statistically significant differences. Outliers indicate that both segments do not share the same most recent common ancestor, and therefore differences are due to reassortment. Additional control measures must be considered to guard against the potential for selective pressures to skew results. Rabadan et al., attempted to get around this problem examining only synonymous mutations. Applying this technique, the authors predict a lower bound on reassortment

frequency for seasonal influenza circulating in New York and New Zealand for the years between 1999 through 2006 to be between 2-3 reassortments a years. This method as well as others has the potential – with significant additional effort to apply beyond reassortment to detecting recombination in non-segmented viruses and in theory bacteria. We are currently in the process of obtaining a new reassortment detection software package from the Center for Bioinformatics and Computational Biology at the University of Maryland to determine its potential to be incorporated into a viral shift detection sequence analysis pipeline. The larger sizes of bacterial genomes relative to viruses will pose significant additional challenges to the design of an efficient recombination detection pipeline.

Once the appropriate reassortment detection method is determined, we can begin to characterize the relationship between reassortment and single point mutations. Does reassortment affect the mutational dynamics of the influenza genome in a way that would further alter patterns of genetic drift? An important question for the rapid design of new medical counter measures could depend on whether these new forms of genetic information can be used to identify markers associated with (and ideally drivers of) pandemic potential in influenza and other viruses. A recent example of this type of effort is given in Allen et al., 2009.

Bacterial evolution via genetic drift and shift

Introduction to bacterial evolution

Like viruses, bacteria evolve by both genetic drift and genetic shift. Bacteria have mutation frequencies of about 10^{-8} to 10^{-9}; much lower rates as compared to RNA viruses. This lower rate of mutation allows the bacterial genome to be much larger than that of RNA viruses which have a mutation rate of 10^{-3}-10^{-5} substitutions per nucleotide replication. Interestingly, naturally-occurring hypermutable bacterial strains which have mutation frequencies 10-1000 higher than that of wild type strains are isolated more frequently from clinical environments as compared to other environments (Hall and Henderson-Begg, 2006). This increased mutation rate is due to mutations in genes that function in the repair of mutations. Hypermutability appears to be deleterious in vitro but advantageous under certain types of selective pressure (for example, host immune response or antibiotic treatment).

Genetic shift occurs in bacteria by horizontal gene transfer (HGT) which includes the addition of genes or genomic islands (GIs) via conjugation (i.e. conjugative plasmids), transduction (movement of gene segments by bacteriophage), or in some cases, transformation (naturally occurring uptake of DNA from the environment) (see review by Juhas et al., 2009). Species of bacteria differ in the impact that each of these types of HGT has on their evolution, with HGT being much more common in some species of bacteria than others. For example, some species of

bacteria take up DNA from the environment more readily than others. The rate of interchromosomal and intrachromosomal recombination also differs between bacterial species.

Antibiotic resistance can occur via genetic drift or genetic shift (Martinez, 2009). The environment occupied by the bacterial species also affects the propensity for genetic shift to occur. For example, water contaminated by human activities frequently contains antibiotics and antibiotic resistance genes which increase the likelihood of bacteria present in these environments becoming resistant to various antibiotics. Multidrug resistant bacteria are also present in soils, and many species have yet to be identified as they are not culturable in the laboratory. Data from functional screens and metagenomic studies indicate that there may be a high density of resistance genes present in a diversity of sample types (D'Costa et al., 2007, Sommer et al., 2009).

The use of comprehensive methods such as metagenomics and microarray analysis to study of bacterial communities in soils, surface water, and even human microbial flora (i.e. human gut microbiome) samples is critical for detection of novel resistance mechanisms prior to emergence of the mechanism in a clinical setting. As described by D'Costa et al. (2006), "The soil could thus serve as an underrecognized reservoir for resistance that has already emerged or has the potential to emerge in clinically important bacteria. Consequently, an understanding of resistance determinants present in the soil—the soil resistome—will provide information not only about antibiotic resistance frequencies but also about new mechanisms that may emerge as clinical problems." A whole genome tiling microarray for detection of nucleotide sequences important for antibiotic resistance for biothreat bacterial organisms is currently under development and testing at LLNL as part of a funded DHS effort.

Overview of the role of genetic shift and drift in the evolution of bacterial select agents

The *Burkholderia pseudomallei* genome consists of two chromosomes and no plasmids. The large chromosome primarily contains genes essential for growth, and the small chromosome contains more diverse genes that are primarily involved in exploiting variable environmental conditions. HGT plays a central role in the evolution of *B. pseudomallei* (Figure 3). Whole genome sequencing of five *B. pseudomallei* isolates revealed the presence of 71 distinct GIs ranging in size from 4 to 100 kb. Strain differences in GI content may affect virulence levels (Tuanyok et al., 2008). For example, acquisition of a GI that enables capsule formation increased ability to evade immune response, acquisition of a GI encoding fimbriae-associated genes increased adherence to host tissues, and gene loss which indirectly led to increased TSS (type III secretion system) activity and increased virulence.

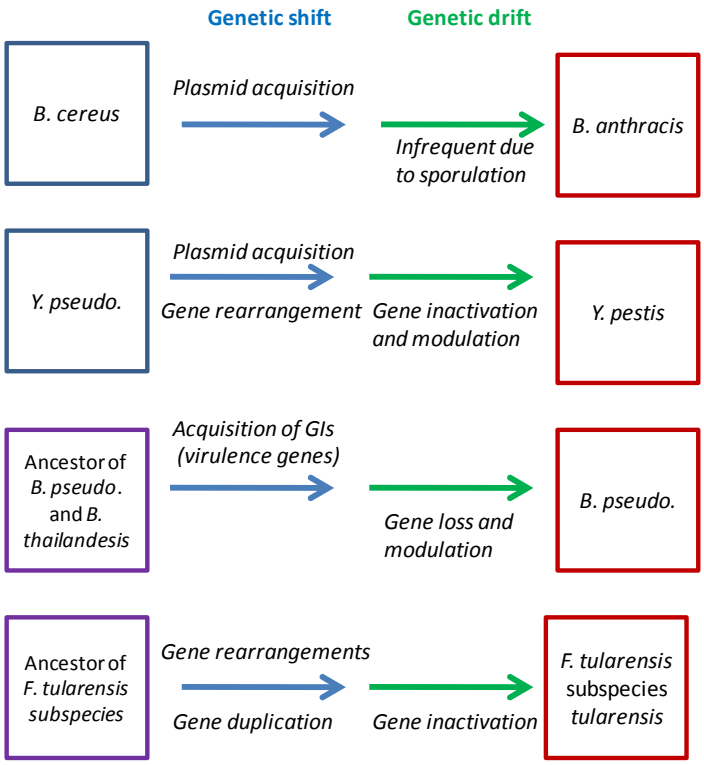

Figure 3. The role of genetic shift and drift in the evolution of bacterial select agents.

Yersinia pestis is believed to be derived from a clone of *Y. pseudotuberculosis* which evolved first by genetic shift (gain of plasmids and intrachromosomal recombination), and then by genetic drift (sequence polymorphisms differentiate genomes of *Y. pestis* strains) (Achtman and Wagner 2008; Chain et al., 2004). The acquisition of two plasmids (pMT1, pCP1) play a key role in flea-borne transmission and thus survival of *Y. pestis* in the environment. Persistent infection of the flea may also provide an opportunity for the *Y. pestis* genome to acquire genes from other bacteria present in the gut of the flea (Galimand et al., 2006). Although antibiotic resistance is rarely detected in *Y. pestis* strains, in the two cases of human plague where resistant strains were identified, the genes responsible for the resistance were present on unique plasmids.

B. anthracis appears to be derived from the clonal expansion of a single ancestral *B. cereus* strain that acquired the two virulence plasmids (pXO1 and pXO2) and the nonsense *plcR* mutation. Genetic drift occurs extremely slowly in *B. anthracis* strains due in large part to the life style of the bacterium; sporulation precludes genetic mutation. A detailed analysis of genetic drift in *B. anthracis* is in the section below. A review of the literature reveals no large differences in antibiotic resistance between isolates of *B. anthracis* (Coker et al., 2007).

The *F. tularensis* subspecies *tularensis*, *holarctica* and *mediasiatica* have not been found to contain plasmids. In contrast, the *novicida* subspecies has been shown to contain plasmid pFNL10.The *Francisella tularensis* subspecies *tularensis* genome is characterized by the presence of many insertion elements leading to genomic rearrangements, gene inactivation in the form of pseudogenes (mainly due to the insertion elements, point mutations, and small indels), and gene duplications which may lead to increased virulence (Rohmer et al., 2007; Larsson et al., 2009).

Finding evidence for genetic drift in *Bacillus anthracis*

LLNL has access to data from bacterial samples grown in laboratory experiments using neutral growth conditions where the genome sequence is available for the initial sample and a descendant sample after 100, 000 cell doublings of growth. The largest of the available datasets, seven distinct *Bacillus anthracis* strains and their respective laboratory grown descendants, were examined to consider the role genetic drift might play in the evolution of *B. anthracis* strains. One of the key initial observations was the large number of mutations between parent and descendant, which could not be distinguished from potential sequencing errors. Each of the sequenced lab strains were only completed to "draft" quality using traditional Sanger sequencing making the potential for sequencing error high and thus limiting much of the data's value for genetic variation analysis. This issue must be taken into careful consideration prior to the initiation of future studies to ensure that sufficient sequencing resources are dedicated to producing the low sequencing error rates needed for reliable genetic polymorphism detection. A potential guide in this regard using second-generation sequencing technology may come from a recently published more elaborate study on bacterial evolution for *Escherichia coli* where Illumina sequencing was used with a minimum coverage of 55x per genome position along with the use of additional quality controls (Barrick et al, 2009).

To control for sequencing error in the existing datasets, a stop gap measure was employed. When a single nucleotide polymorphism was identified between the parent strain and its descendant and the mutation could be found in a second *B. anthracis* genome, the mutation was counted as a true genetic mutation. (All available non-redundant *B. anthracis* genome data was collectively aligned.) In order to avoid false polymorphism detection due to ambiguity in the alignments, an added quality control measure required that polymorphisms must be surrounded on the upstream and downstream ends by at least 4 exact matching nucleotides between the passaged *B.anthracis* pairs. A total of 28 nucleotide substitutions mapped to 16 distinct genome locations of which 15 mapped directly to annotations in the reference *B. anthracis* Ames strain (with one mapping to a deleted region of Ames). Results are summarized in Table 2.

Table 2. *B. anthracis* strains grown in the laboratory.

***B. anthracis* strain (Ba)**	**Total Mutations**	**Tn/Tv***
A0193	8	8/0
A0488	3	3/0
A0442	6	5/1
A0248	3	3/0
A0465	2	2/0
A0389	3	3/0
A0174	3	3/0

*Tn=Transition/Tv=Transversion count
(Transition represents a Purine to Purine or Pyrimidine to Pyrimidine mutation andTransversion represents a mutation between Purine and Pyrimidine)

Of the 15 mutations mapped to the Ames annotation, 13 mapped to genes of which 9 were synonymous mutations. Results are summarized in Table 3. Given that there are more candidate mutations within a gene that would lead to non-synonymous changes, if replication errors occur randomly across the genome more non-synonymous mutations would be expected. Since 9 of the 13 gene mutations are synonymous, negative selective pressure likely plays an important role in the evolution of the descendant. The "From/To" column shows the number of *B. anthracis* strains that share the parental variant's mutation versus the number of *B. anthracis* strains that share the lab passaged strain's mutation. There were a total of 32 sequences included in the analysis so that when the "From" + "To" does not equal 32, it means that mutations did not map to each of the sequences. Since many of the sequences are draft quality with sequences missing this is expected. For non-synonymous mutations, there is a clear pattern of going from the minority variant in the ancestral strain to acquiring the dominant variant. However, there is no such clear trend showing up in the synonymous mutation sites, supporting the notion that these sites are free of selective pressure and may continue to mutate throughout the population with no clear persistent dominant version.

Table 3. Mutations in Genes. The table shows gene description, strain with mutation, amino acid variants, GenBank identifier, count of strains sharing ancestor or descendant mutant (From/To) and mutant position in the Ames genome.

Gene	Strain	Mutation	GenBank ID	From/To	Position
Protein with cell wall anchor domain	A0488	I->M	NP_842941	7/24	416200
	A0488	G->E	NP_842941	7/24	416212
Hypothetical Protein	A0442	N->S	NP_843925	6/25	1390278
Hypothetical Protein	A0248	A->T	NP_844826	8/16	2280723
Hypothetical Protein	A0442	V	NP_845276	22/4	2706909
Spore coat assembly protein SafA	A0193	G	NP_846874	9/11	4234049
	A0193	N	NP_846874	9/11	4234055
	A0465, A0248, A0389, A0442	G	NP_846874	19/9	4234088
	A0465, A0248, A0389, A0193, A0442	N	NP_846874	18/10	4234094
	A0193	V	NP_846874	14/14	4234106
	A0193, A0174	N	NP_846874	2/29	4234133
	A0193, A0174, A0442	V	NP_846874	17/14	4234145
Hypothetical Protein	A0389	A	NP_846970	8/24	4332538

The fact that 5 of the 7 parent/descendant pairs show evidence for their own distinct mutations indicates that genetic drift could very well contribute to the differences observed in the laboratory strains. The three strains with unique non-synonymous mutations show the potential for the laboratory strain to acquire phenotypic changes due to drift. However, it cannot be ruled out that each of these strain specific mutations confers a selective advantage that is unique to a specific strain. Ideally, we would like to look at data from multiple independent passage

experiments beginning with the same parental strain to strengthen support for the claim of drift in these cases. The concentration of mutations at the *safA* locus shows a potentially non-random concentration of synonymous mutations within the genome. The SafA protein plays a critical role in the assembly of the spore coat and mutations in the *safA* gene makes the spore susceptible to lysozyme inactivation by the host immune defenses, and decreases the resistance of spores to digestion by soil nematodes (Laaberki and Dworkin, 2008). However, because sporulation is not critical to passage of vegetative cells in the laboratory, mutations in the *safA* gene may not be disadvantageous (Lai et al., 2003).

The concentration of mutations at a gene locus could also have important implications for predicting the erosion of DNA signatures used in broad range diagnostics. However, the number of observed mutations still remains small enough to suggest that any diagnostic signature set would not be substantially impacted by short term drift. While it is tempting to extrapolate further on the implications to TMTI's medical diagnostic and countermeasure objectives, it is not possible to draw broader conclusions on the observed drift patterns without more sampled time points and higher quality sequence data. Higher quality sequence data and biological experimentation is also needed to attribute any biological significance to the concentration of mutations in the *safA* gene. We have identified that the original sample material from the passage experiments is still available and could be re-sequenced with newer sequencing technology to provide a more complete picture of how *Bacillus anthracis* evolves under laboratory conditions and how this form of evolution compares to the inferred evolutionary process of the wild type species.

Data Gaps and potential solutions

Ideally, the role of genetic drift and shift in pathogen evolution is best addressed by sequencing of epidemiologically-linked, longitudinal samples that have not been passaged in the laboratory and have been collected from natural infections or outbreaks. In the case of RNA viruses, it is important to employ ultra-deep sequencing techniques to identify the role of quasispecies in viral evolution. In the case of select agent viruses, such sample collections are generally not available. However, longitudinally collected tissue samples from a rabies virus outbreak (and species jump) are available and will be analyzed using ultra-deep sequence analysis. As the rabies virus is a zoonotic virus with a genome that is similar in structure to filoviruses, this information should provide important insights into the role of genetic drift in interspecies transmission. Similarly, work has recently been funded to use whole genome sequencing to examine the role of genetic shift (recombination) in adaptation of a zoonotic coronavirus to human cells.

As discussed in the section on influenza evolution, it is increasingly apparent that genetic shift via reassortment is an important mechanism of evolution for segmented viruses. Therefore, detection assays should be designed that include reagents for detection of each genomic segment

so reassortant viruses are recognized, and computational software should be designed that allow rapid detection of reassortment and recombination.

Many of the select agent viruses and bacteria are vectored by arthropods including Crimean-Congo hemorrhagic fever virus, Rift Valley fever virus, eastern equine encephalitis virus, Venezuelan equine encephalitis virus, *Francisella tularensis*, and *Yersinia pestis*. Although the reliance on an arthropod vector generally acts to constrain genetic drift, infection of the arthropod vector also provides an opportunity for genetic shift (reassortment, recombination and lateral gene transfer) to occur. Therefore, understanding of the microbial flora of vectors may provide insight into reservoirs of genetic material available to arthropod-borne pathogens. For example, a group of flaviviruses have recently been found to be harbored as gut flora by various species of Culex mosquitoes (Kim et al., 2009). Although these viruses do not appear to be human pathogens, other flaviviruses such as yellow fever virus do cause human disease, and the role of co-infecting mosquito flaviviruses in the evolution of the human pathogens has not been determined. Collaboration with Bob Tesh and Scott Weaver at University of Texas, Medical Branch may provide insight into the significance of arthropod flora in the evolution of arthropod-borne human pathogens.

The role of environmental microbes in lateral gene transfer of genomic islands and antibiotic resistance genes is poorly understood. Metagenomic analysis of the natural environments of select agent pathogens may provide insights into the genetic diversity present in these environments. For example, metagenomic analysis of soils from *Burkholderia pseudomallei* endemic regions may identify the reservoirs for the genomic islands present in clinical isolates and identify novel genomic islands that may be transferred to *Burkholderia pseudomallei*. Additionally, the role of nematodes in the life cycle of human pathogens such as *Francisella tularensis* and *Yersinia pestis* remains to be determined. However, processing of metagenomic samples presents a challenge due the presence of eukaryotic nucleic acids as well as inhibitory substances. The use of microarrays for enrichment of bacterial or viral nucleic acid combined with next generation sequence technology and advance computational tools may allow these challenges to be addressed.

References

Achtman, M. and Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. Nat. Rev. Microbiol. *6*, 431-440.

Allen, J.E., Gardner S.N., Vitalis E.A., Slezak T.R. (2009). Conserved amino acid markers from past influenza pandemic strains. BMC Microbiology 9:77.

Babkin, I.V. and Shchelkunov, S.N. (2006). The time scale in poxvirus evolution. Mol. Biol. *40*, 20-24.

Babkin, I.V., Nepomniashchikh, T.S., Maksiutov, R.A., Gutorov, V.V., Babkina, I.N., and Shchelkunov, S.N. (2008). Comparative analysis of variable regions in the genomes of variola virus. Mol. Biol. *42* , 612-624.

Barrick, J.E., Yu, D.S., Yoon, S.H., Jeong, H., Oh, T.K., Schneider, D., Lenski, R.E., and Kim, J.F. (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature *461*, 1243-1247.

Bird, B.H., Khristova, M.L., Rollin, P.E., Ksiazek, T.G., and Nichol, S.T. (2007). Complete genome analysis of 33 ecologically and biologically diverse Rift Valley fever virus strains reveals widespread virus movement and low genetic diversity due to recent common ancestry. J. Virol. *81*, 2805-2816.

Boni, M.F., Zhou, Y., Taubenberger, J.K., and Holmes, E.C. (2008). Homologous recombination is very rare or absent in human influenza A virus. J. Virol. *82*, 4807-4811.

Borucki, M.K., Chandler, L.J., Parker, B.M., Blair, C.D., and Beaty, B.J. (1999). Bunyavirus superinfection and segment reassortment in transovarially infected mosquitoes. J. Gen. Virol. *80 (Pt 12)*, 3173-3179.

Briese, T., Kapoor, V., and Lipkin, W.I. (2007). Natural M-segment reassortment in Potosi and Main Drain viruses: implications for the evolution of orthobunyaviruses. Arch. Virol. *152*, 2237-2247.

Chain, P.S., Carniel, E., Larimer, F.W., Lamerdin, J., Stoutland, P.O., Regala, W.M., Georgescu, A.M., Vergez, L.M., Land, M.L., Motin, V.L., Brubaker, R.R., Fowler, J., Hinnebusch, J., Marceau, M., Medigue, C., Simonet, M., Chenal-Francisque, V., Souza, B., Dacheux, D., Elliott, J.M., Derbise, A., Hauser, L.J., and Garcia, E. (2004). Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis.* Proc. Natl. Acad. Sci. U. S. A *101*, 13826-13831.

Chare, E.R., Gould, E.A., and Holmes, E.C. (2003). Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. J. Gen. Virol. *84*, 2691-2703.

Coker, P.R., Smith, K.L., and Hugh-Jones, M.E. (2002). Antimicrobial susceptibilities of diverse *Bacillus anthracis* isolates. Antimicrob. Agents Chemother. *46,* 3843-3845.

D'Costa, V.M., Griffiths, E., and Wright, G.D. (2007). Expanding the soil antibiotic resistome: exploring environmental diversity. Curr. Opin. Microbiol. *10*, 481-489.

Deyde, V.M., Khristova, M.L., Rollin, P.E., Ksiazek, T.G., and Nichol, S.T. (2006). Crimean-Congo hemorrhagic fever virus genomics and global diversity. J. Virol. *80*, 8834-8842.

Domingo, E. (2007). Virus Evolution. In Fields Virology, D.M.Knipe and P.M.Howley, eds. (Philadelphia: Lippincott Williams & Wilkins), pp. 389-421.

Elena, S.F., Carrasco, P., Daros, J.A., and Sanjuan, R. (2006). Mechanisms of genetic robustness in RNA viruses. EMBO Rep. *7*, 168-173.

Emonet, S.F., de la Torre, J.C., Domingo, E., and Sevilla, N. (2009). Arenavirus genetic diversity and its biological implications. Infect. Genet. Evol. *9*, 417-429.

Esposito, J.J., Sammons, S.A., Frace, A.M., Osborne, J.D., Olsen-Rasmussen, M., Zhang, M., Govil, D., Damon, I.K., Kline, R., Laker, M., Li, Y., Smith, G.L., Meyer, H., Leduc, J.W., and Wohlhueter, R.M. (2006). Genome sequence diversity and clues to the evolution of variola (smallpox) virus. Science *313*, 807-812.

Galimand, M., Carniel, E., and Courvalin, P. (2006). Resistance of *Yersinia pestis* to antimicrobial agents. Antimicrob. Agents Chemother. *50*, 3233-3236.

Gerrard, S.R., Li, L., Barrett, A.D., and Nichol, S.T. (2004). Ngari virus is a Bunyamwera virus reassortant that can be associated with large outbreaks of hemorrhagic fever in Africa. J. Virol. *78*, 8922-8926.

Ghedin, E., Sengamalay, N.A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D.J., Sitz, J., Koo, H., Bolotov, P., Dernovoy, D., Tatusova, T., Bao, Y., St, G.K., Taylor, J., Lipman, D.J., Fraser, C.M., Taubenberger, J.K., and Salzberg, S.L. (2005). Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. Nature *437*, 1162-1166.

Gubler D.J., Kuno, G., and Markoff, L. (2007). Flaviviruses. In Fields Virology, D.M.Knipe and P.M.Howley, eds. (Philadelphia: Lippincott Williams & Wilkins), pp. 1184-1190.

Hall, L.M. and Henderson-Begg, S.K. (2006). Hypermutable bacteria isolated from humans--a critical analysis. Microbiology *152*, 2505-2514.

Hanada, K., Suzuki, Y., and Gojobori, T. (2004). A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol. Biol. Evol. *21*, 1074-1080.

Holmes, E.C., Ghedin, E., Miller, N., Taylor, J., Bao, Y., St, G.K., Grenfell, B.T., Salzberg, S.L., Fraser, C.M., Lipman, D.J., and Taubenberger, J.K. (2005). Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. PLoS. Biol. *3*, e300.

Huson, D.H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. *23*, 254-267.

Juhas, M., van, d.M., Jr., Gaillard, M., Harding, R.M., Hood, D.W., and Crook, D.W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol. Rev. *33*, 376-393.

Kim, D.Y., Guzman,H., Bueno,R., Jr., Dennett,J.A., Auguste,A.J., Carrington,C.V., Popov,V.L., Weaver,S.C., Beasley,D.W., and Tesh,R.B. (2009). Characterization of Culex Flavivirus (Flaviviridae) strains isolated from mosquitoes in the United States and Trinidad. Virology *386*, 154-159.

Laaberki, M.H. and Dworkin, J. (2008). Role of spore coat proteins in the resistance of Bacillus subtilis spores to Caenorhabditis elegans predation. J. Bacteriol. *190*, 6197-6203.

Lai, E.M., Phadke, N.D., Kachman, M.T., Giorno, R., Vazquez, S., Vazquez, J.A., Maddock, J.R., and Driks, A. (2003). Proteomic analysis of the spore coats of *Bacillus subtilis* and *Bacillus anthracis*. J. Bacteriol. *185*, 1443-1454.

Lai, M., Perlman, S., and Anderson, L.J. (2007). Coronaviridae. In Fields Virology, D.M.Knipe and P.M.Howley, eds. (Philadelphia: Lippincott Williams & Wilkins), pp. 1305-1336.

Larsson, P., Elfsmark, D., Svensson, K., Wikstrom, P., Forsman, M., Brettin, T., Keim, P., and Johansson, A. (2009). Molecular evolutionary consequences of niche restriction in *Francisella tularensis*, a facultative intracellular pathogen. PLoS. Pathog. *5*, e1000472.

Li, W., Sui, J., Huang, I.C., Kuhn, J.H., Radoshitzky, S.R., Marasco, W.A., Choe, H., and Farzan, M. (2007). The S proteins of human coronavirus NL63 and severe acute respiratory syndrome coronavirus bind overlapping regions of ACE2. Virology *367*, 367-374.

Martinez, J.L. (2009). The role of natural environments in the evolution of resistance traits in pathogenic bacteria. Proc. Biol. Sci. *276*, 2521-2530.

Moya, A., Holmes, E.C., and Gonzalez-Candelas, F. (2004). The population genetics and evolutionary epidemiology of RNA viruses. Nat. Rev. Microbiol. *2*, 279-288.

Nagarajan, N and C Kingsford, C. (2008).Uncovering Reassortments Among influenza Strains by Enumerating Maximal Bicliques. In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, pages 223-230.

Nakhleh, L., Warnow, T., Linder, C.R., and St, J.K. (2005). Reconstructing reticulate evolution in species-theory and practice. J. Comput. Biol. *12*, 796-811.

Olsen, C.W., Karasin, A.I., Carman, S., Li, Y., Bastien, N., Ojkic, D., Alves, D., Charbonneau, G., Henning, B.M., Low, D.E., Burton, L., and Broukhanski, G. (2006). Triple reassortant H3N2 influenza A viruses, Canada, 2005. Emerg. Infect. Dis. *12*, 1132-1135.

Parrish, C.R., Holmes, E.C., Morens, D.M., Park, E.C., Burke, D.S., Calisher, C.H., Laughlin, C.A., Saif, L.J., and Daszak, P. (2008). Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol. Mol. Biol. Rev. *72*, 457-470.

Pourrut, X., Delicat, A., Rollin, P.E., Ksiazek, T.G., Gonzalez, J.P., and Leroy, E.M. (2007). Spatial and temporal patterns of Zaire ebolavirus antibody prevalence in the possible reservoir bat species. J. Infect. Dis. *196 Suppl 2*, S176-S183.

Rabadan, R., Levine, A.J., and Krasnitz, M. (2008). Non-random reassortment in human influenza A viruses. Influenza. Other Respi. Viruses. *2*, 9-22.

Reese, S.M., Blitvich, B.J., Blair, C.D., Geske, D., Beaty, B.J., and Black, W.C. (2008). Potential for La Crosse virus segment reassortment in nature. Virology Journal, *5*, 164.

Rohmer, L., Fong, C., Abmayr, S., Wasnick, M., Larson Freeman, T.J., Radey, M., Guina, T., Svensson, K., Hayden, H.S., Jacobs, M., Gallagher, L.A., Manoil, C., Ernst, R.K., Drees, B., Buckley, D., Haugen, E., Bovee, D., Zhou, Y., Chang, J., Levy, R., Lim, R., Gillett, W., Guenthener, D., Kang, A., Shaffer, S.A., Taylor, G., Chen, J., Gallis, B., D'Argenio, D.A., Forsman, M., Olson, M.V., Goodlett, D.R., Kaul, R., Miller, S.I., and Brittnacher, M.J. (2007). Comparison of *Francisella tularensis* genomes reveals evolutionary events associated with the emergence of human pathogenic strains. Genome Biol. *8*, R102.

Sharp, P.M. (2002). Origins of human virus diversity. Cell *108*, 305-312.

Simmonds, P. (2004). Genetic diversity and evolution of hepatitis C virus--15 years on. J. Gen. Virol. *85*, 3173-3188.

Sommer, M.O., Dantas, G., and Church, G.M. (2009). Functional characterization of the antibiotic resistance reservoir in the human microflora. Science *325*, 1128-1131.

Tuanyok, A., Leadem, B.R., Auerbach, R.K., Beckstrom-Sternberg, S.M., Beckstrom-Sternberg, J.S., Mayo, M., Wuthiekanun, V., Brettin, T.S., Nierman, W.C., Peacock, S.J., Currie, B.J., Wagner, D.M., and Keim, P. (2008). Genomic islands from five strains of *Burkholderia pseudomallei*. BMC. Genomics *9*, 566.

Vasilakis, N., Deardorff, E.R., Kenney, J.L., Rossi, S.L., Hanley, K.A., and Weaver, S.C. (2009). Mosquitoes put the brake on arbovirus evolution: experimental evolution reveals slower mutation accumulation in mosquito than vertebrate cells. PLoS. Pathog. *5*, e1000467.

Wittmann, T.J., Biek, R., Hassanin, A., Rouquet, P., Reed, P., Yaba, P., Pourrut, X., Real, L.A., Gonzalez, J.P., and Leroy, E.M. (2007). Isolates of Zaire ebolavirus from wild apes reveal genetic lineage and recombinants. Proc. Natl. Acad. Sci. U. S. A *104*, 17123-17127.

Woolhouse, M. and Gaunt, E. (2007). Ecological origins of novel human pathogens. Crit Rev. Microbiol. *33*, 231-242.